



Processing NASA Earth Science Data on Nebula Cloud

Aijun Chen^{1,2}, Long Pham¹, and Steven Kempler¹

Dr. AIJUN CHEN

¹Goddard Earth Sciences Data and Information Services Center (GES DISC)
NASA Goddard Space Flight Center (GSFC)

²Center for Spatial Information Science and Systems (CSISS)
George Mason University (GMU)



Goddard Earth Science DISC



❖ NASA GES DISC offers atmospheric related observation and model data and applied services.

❖ Data in missions:

- TRMM (PR, TMI, VIRS),
- Terra (MODIS, ASTER),
- Aqua (AIRS, MODIS, AMSU-A, HSB),
- Aura (MLS, HIRDLS, OMI, TES),
- CloudSat,
- CALIPSO, etc.

❖ Services and Tools:

- **Mirador**,
- **Giovanni**,
- OPeNDAP,
- GrADS,
- OGC WMS,
- FTP, etc.



Goddard Earth Science DISC - Mirador



**National Aeronautics
and Space Administration**

Goddard Earth Sciences
Data and Information Services Center

Search DISC

+ Advanced Search + 60

+ ATMOS COMPOSITION

+ HYDROLOGY

+ A-TRAIN

+ AIRS

+ MODELING

+ MAIRS

+ MEASURES

+ PRECIPITATION

+ GES DISC Home

Mirador

+ OVERVIEW

+ HELP CENTER

+ DATA HOLDINGS

+ VIEW CART

Additional Features

+ News

+ Restricted Data

+ Feedback

+ FAQ

Mirador
Data Access Made Simple

You are here: [Keyword Search](#)

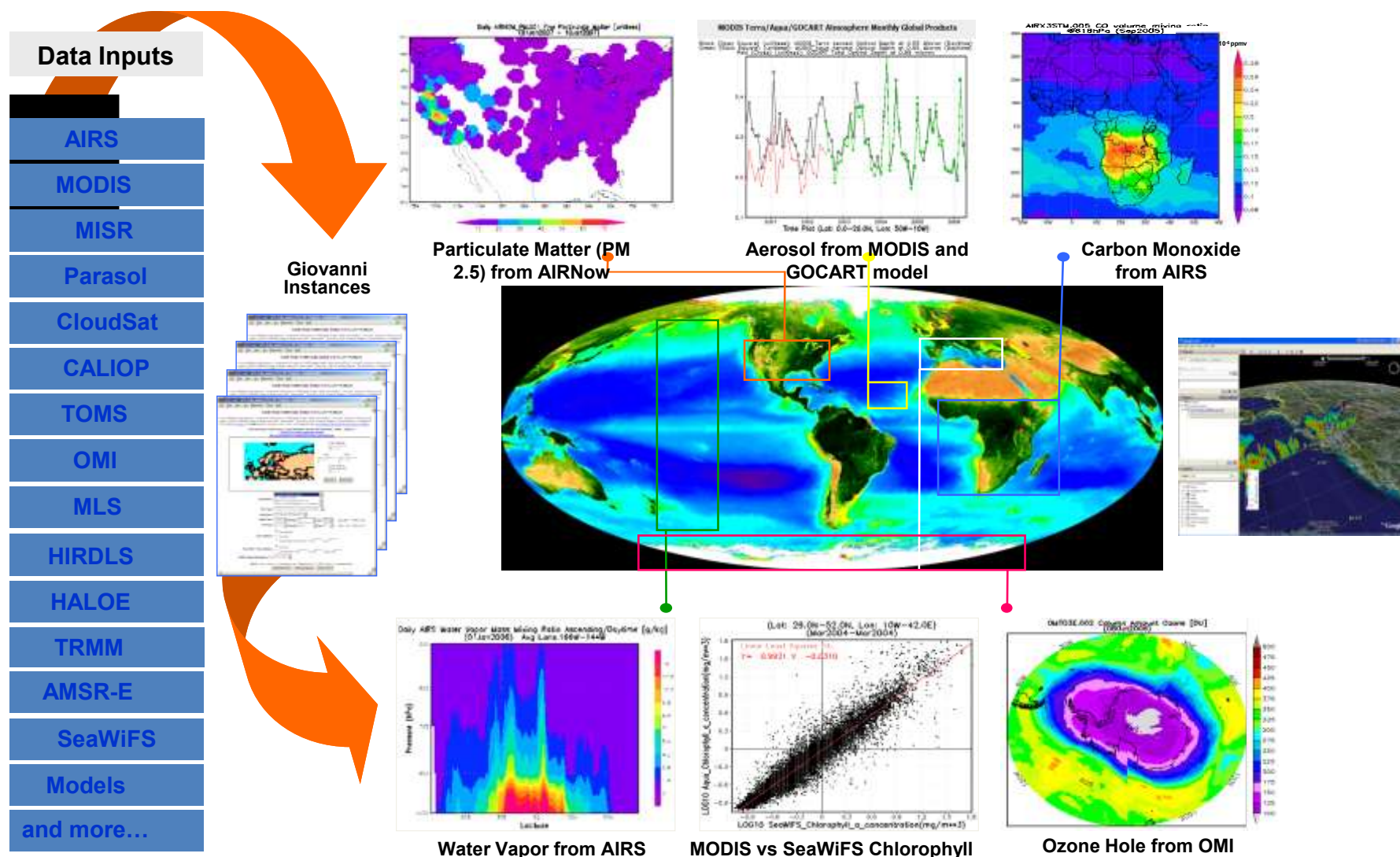
[Keyword](#) [Projects](#) [Science Areas](#)

[Keyword:](#) [Time Span:](#)
[Location:](#) [Update Map](#) [Search GES-DISC](#)





Goddard Earth Science DISC -Giovanni

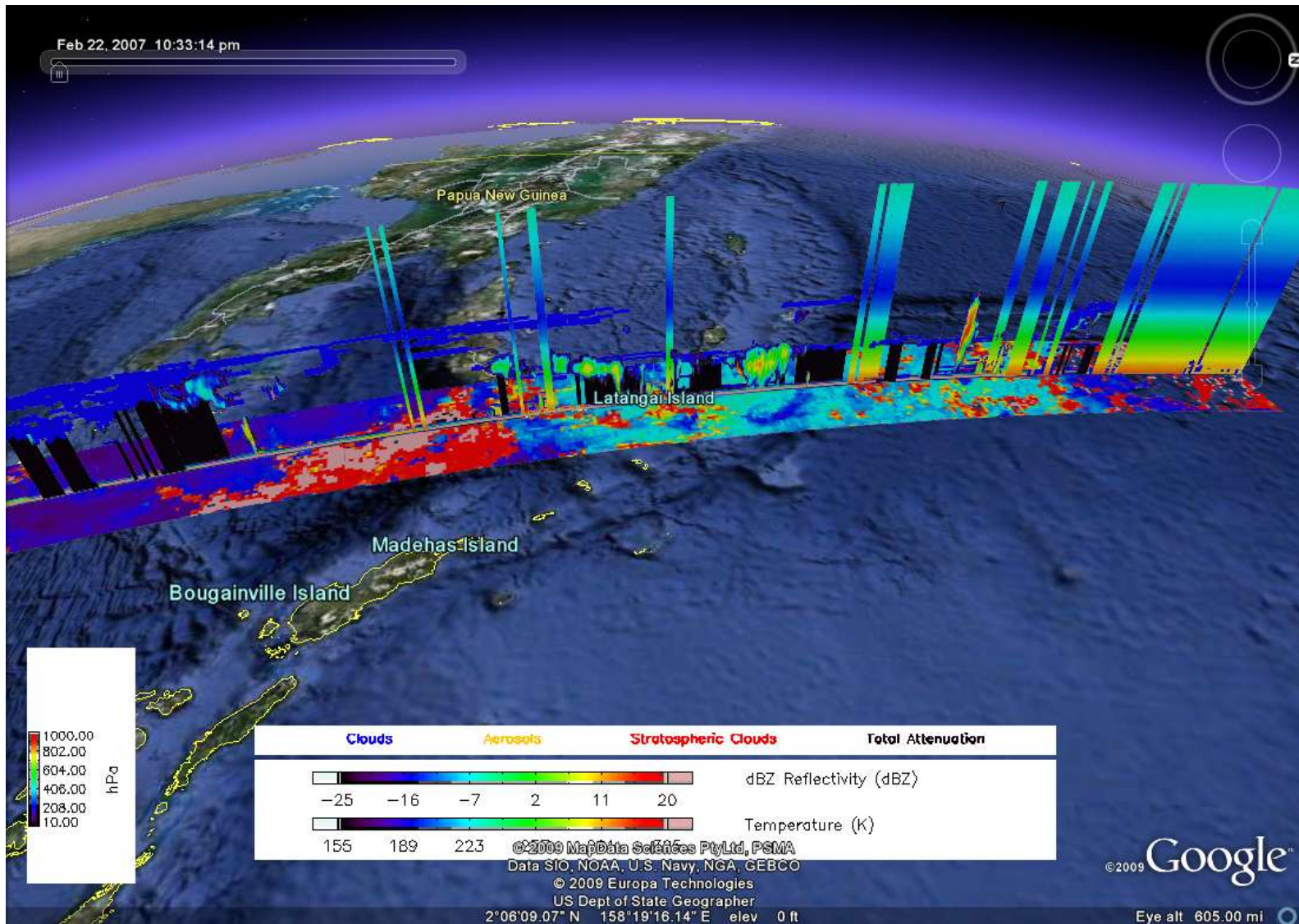


Courtesy of Suhung Shen, NASA GES DISC

NASA/GMU CSISS



Goddard Earth Science DISC - Visualization





Goddard Earth Science DISC



<http://disc.gsfc.nasa.gov>



Outline



- Background
- The migrating procedure
- Performance estimation, comparison, and analysis
- Costs estimation and comparison
- Output verification
- Advantages of Nebula Cloud
- Challenges and Lessons Learned
- Summary



Background -1



- ❖ Cloud Computing has been implemented and used by quite a few commercial companies (e.g. Amazon EC2 [SaaS, 2006], Google App Engine [PaaS, 2008], Microsoft Azure [PaaS, 2008], etc.).
- ❖ NASA Launched Nebula in 2008 to provide Infrastructure as a Service (IaaS).
 - a) Make NASA realize significant cost savings through efficient resource utilization, reduced energy consumption, and reduced labor costs.
 - b) Provide an easier way for NASA scientists and researchers to efficiently explore and share large and complex data sets.
 - c) Allow customers to provision, manage, and decommission computing capabilities on an as-needed basis.



NASA Nebula: <http://nebula.nasa.gov/>



Background -2

- ❖ GES DISC has been evaluating feasibility and suitability of migrating GES DISC's applications to the Nebula platform by porting following projects.
 - a) Using Nebula Cloud to run scientific data processing infrastructure**
S4PM is an open source data processing infrastructure. Based on S4PM, scientific data processing algorithms can be run to efficiently process large volumes of satellite data. <http://sourceforge.net/projects/s4pm/>
 - b) Using Nebula Cloud to run scientific data processing workflow**
The Atmospheric Infrared Sounder (AIRS) focuses on supporting climate research and improving weather forecasting. Based on S4PM, the **AIRS Level 1 & Level 2 algorithms workflow**, consisting of many of sub-algorithms (executables), processes large volumes of AIRS Level 0 data to produce Level 1 data as intermediate results, and finally outputs Level 2 data products.



Background -3



c) Porting a Web-based scientific data processing application to Nebula Cloud

Giovanni is a Web-based application which offer online visualization and analysis of vast amounts of Earth science data. The **Giovanni MAPSS** (Multi-sensor Aerosol Products Sampling System) portal focuses on visualizing aerosol relationships among ground-based data and satellite data.

- ❖ The experiences, lessons learned, and tutorials will expedite our future efforts to utilize Nebula/Cloud computing technologies to process Earth Science data.



The diagram illustrates the MAPSS (Multi-sensor Aerosol Products Sampling System) architecture, showing the flow of data from source to visualization.

Data Sources and Initial Processing:

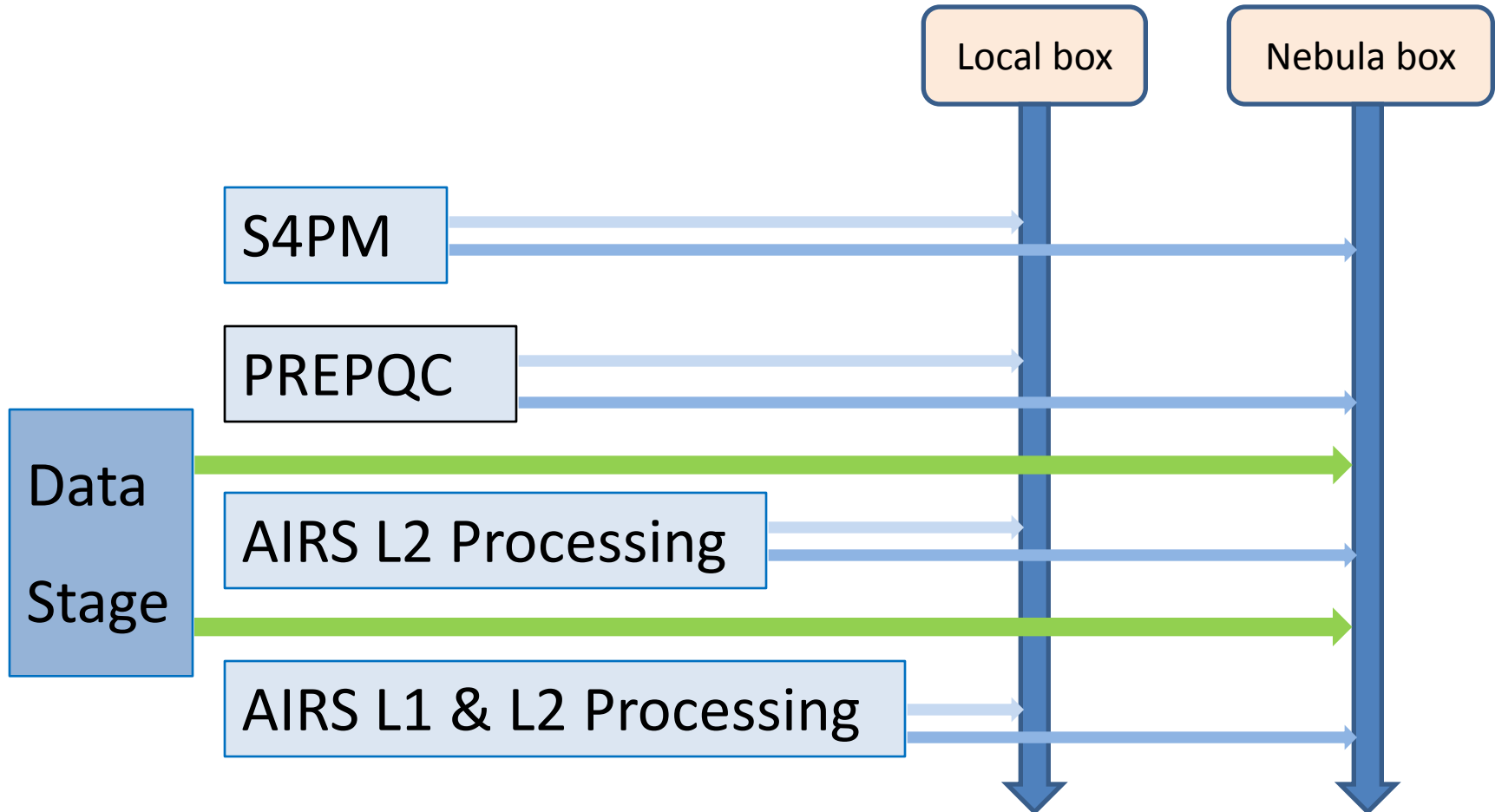
- NASA Nebula Cloud** (Bottom) provides data to **GES DISC S4P/S4PM** and **HDF4/HDF5**.
- GES DISC S4P/S4PM** feeds into **HDF-4.2/SDPTK_5.2.12**.
- HDF-4.2/SDPTK_5.2.12** feeds into **AIRS L1 Processing** and the **AIRS L1/L2 Algorithms workflow**.

Workflow and Visualization:

- The **AIRS L1/L2 Algorithms workflow** feeds into the **Kepler Workflow Engine**.
- The **Kepler Workflow Engine** feeds into **aGiovanni**.
- aGiovanni** feeds into **MAPSS**.
- MAPSS** feeds into the **MAPSS Time Series** visualization (Top Right), which shows a graph of Mean ADD (Mean at ADD at 340km from AIRPORT_ADD_15.8 at COVE) over time (Jan 1 2007 to Oct 4 2007).
- MAPSS** also feeds into the **MAPSS: Multi-sensor Aerosol Products Sampling System** web interface (Top Left), which shows a map of the world with a color-coded overlay.



The Migrating Procedures





Performance Estimation -1

-- Hardware Information

	Local Real Linux box	Nebula virtual Linux Box
Hardware	DELL PowerEdge 6800 with Dual-Core Xeon Processor 7100 series, 4 CPU	DELL PowerEdge c2100 with Quad-Core Xeon Processor 5500 series, 2 CPU per c2100
CPU (GHz)	8 cores * 3.16	16 cores * 2.8
RAM (GB)	16	8
Cache Size (MB)	1	4
Storage	11TB	300GB (200GB in default)
CPU Microarchitecture	65nm NetBurst	45nm Nehalem / 32nm Westmere

Core (65nm) /Penryn (45nm)



Performance Estimation -2

S4PM/GUI for PREPQC, AIRS L1 & L2 Processing

Compose Data Request Tool

Year: 2010 Day of Year: 124 Query Archive

Data Type	HH:	00	02	04	06	08	10	12	14	16	18	20	22
AIR10SCC.001	1	1	1	1	1	1	1	1	1	1	1	1	1
AIR10SCI.001	1	1	1	1	1	1	1	1	1	1	1	1	1
AIR10XNM.001													
AIR20SCI.001	1	1	1	1	1	1	1	1	1	1	1	1	1
AIR20XNM.001													
AIR20XSM.001													
AIRAACAL.001													
AIRASCAL.001													
AIRB0CAH.001	1	1	1	1	1	1	1	1	1	1	1	1	1
AIRB0CAL.001	1	1	1	1	1	1	1	1	1	1	1	1	1
AIRB0CAP.001	1	1	1	1	1	1	1	1	1	1	1	1	1
AIRB0SCI.001	1	1	1	1	1	1	1	1	1	1	1	1	1
AIRH1ENC.001	1	1	1	1	1	1	1	1	1	1	1	1	1
AIRH1ENG.001													
AIRH2ENC.001	1	1	1	1	1	1	1	1	1	1	1	1	1
AIRH2ENG.001													
AVI3_ANH.001	1			1			1				1		
AVI6_ANH.001	1			1			1				1		
AVI9_ANH.001	1			1			1				1		
PM1ATTNR.003	1	1	1	1	1	1	1	1	1	1	1	1	1
PM1EPHND.003							1						
PMCO_HK.001	1	1	1	1	1	1	1	1	1	1	1	1	1

Submit Request Exit

S4PM Monitor: Aqua AIRS A on i-z7900n3r

Station	Jobs	Queue	Max	OK	Fail
Configurator		0	1	0	0
Auto Acquire Airsraw3		0	2	0	0
Auto Acquire Airspar1		0	2	0	0
Auto Acquire Local		0	2	0	0
Auto Acquire Tads1		0	2	0	0
Auto Acquire Airscal2		0	2	0	0
Acquire Data		0	1	43	0
Register Data		0	5	42	0
Select Data		0	80	1698	3
Find Data		0	85	1839	4
Prepare Run		0	5	1838	0
Allocate Disk		0	40	1844	0
Run Algorithm		0	4	1844	0
Register Local Data		0	5	1841	0
Export		0	40	1844	0
Track Data		0	0	3730	0
Receive PAN		0	40	0	0
Sweep Data		0	0	295	0
Repeat Hourly		2	5	0	0
Repeat Daily		4	8	0	0

since 06/03 13:10:59

Refresh Start All Stop All S4PM Admin Tool Algorithm Info Kill All Exit



Performance Comparison -1



--Two-day AIRS L2 Processing at Nebula box and Local box

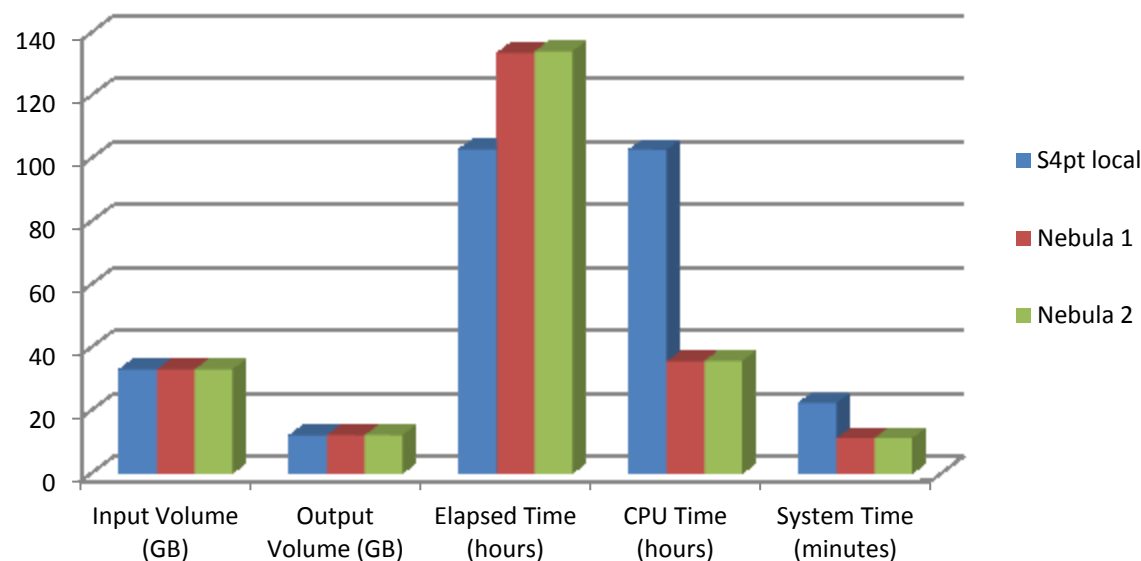
Two days (2010.123-124)	Local Server	Nebula 1	Nebula 2
Input Volume (GB)	33.1	33.1	33.1
Output Volume (GB)	12.16	12.2	12.2
Elapsed Time (hours)	103.05	133.60	134.13
CPU Time (hours)	102.90	35.67	35.80
System Time (minutes)	22.47	11.27	11.27

Input Data (L1):

Calibrated and geolocated radiance in physical units, e.g. brightness temperature in Kelvin (K).

Output Data (L2):

Retrieved physical variables, e.g. temperature, humidity and ozone profiles, total precipitable water, cloud top height.





Performance Comparison -2

--Stable and consistent processing at Nebula box and Local box

AIRS L2 processing at s4pt and Nebula box

One day (2010.123)	Local Server	Nebula 1	Nebula 2
Input Volume L1 data (GB)	15.3	15.3	
Output Volume L2 data (GB)	6.06	6.11	
Elapsed Time (hours)	52.47	17.76	
CPU Time (hours)	52.34	17.76	
System Time (minutes)	10.5	4.34	

Two days (2010.123-124)	Local Server	Nebula 1	Nebula 2
Input Volume L1 data (GB)	33.1	33.1	33.1
Output Volume L2 data (GB)	12.16	12.2	12.2
Elapsed Time (hours)	103.05	133.60	134.13
CPU Time (hours)	102.90	35.67	35.80
System Time (minutes)	22.47	11.27	11.27

Three days (2010.123-125)	Local Server	Nebula 1	Nebula 2
Input Volume L1 data (GB)	48.8	48.8	48.8
Output Volume L2 data (GB)	18.3	18.3	18.3
Elapsed Time (hours)	154.39	207.84	207.83
CPU Time (hours)	154.14	55.31	55.32
System Time (minutes)	32.87	17.23	17.30

Input Data (L1):

Calibrated and geolocated radiance in physical units, e.g. brightness temperature in Kelvin (K).

Output Data (L2):

Retrieved physical variables, e.g. temperature, humidity and ozone profiles, total precipitable water, cloud top height.



Performance Comparison -3

-- Two-day AIRS L1 & L2 Processing at Nebula box and Local box

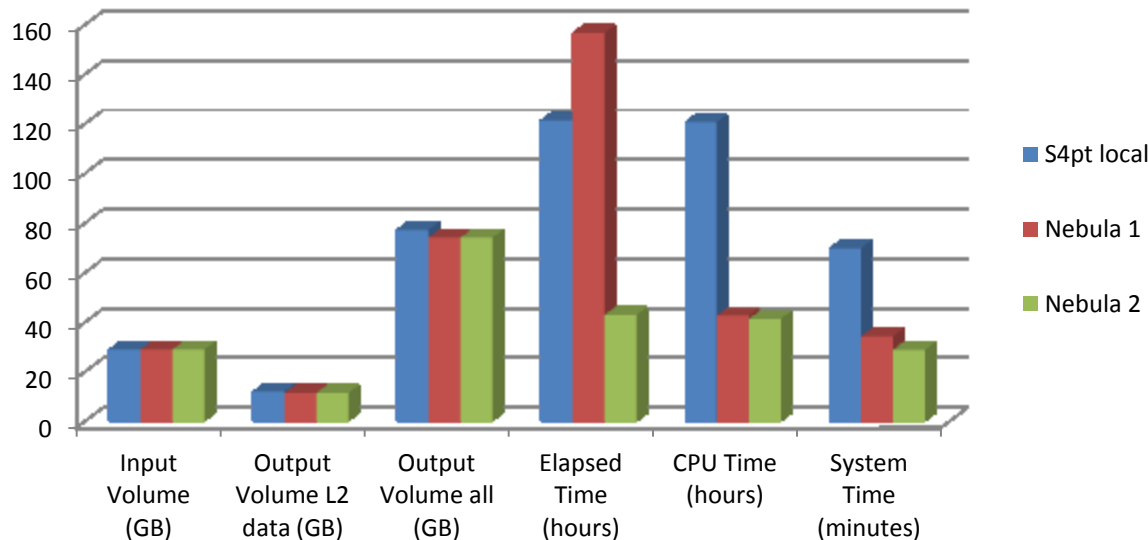
Two days (2010.123-124)	Local Server	Nebula 1	Nebula 2
Input Volume (GB)	29.11	29.11	29.11
Output Volume L2 data (GB)	12.14	11.61	11.64
Output Volume all (GB)	77.47	74.37	74.35
Elapsed Time (hours)	121.70h	157.00h	43.11h
CPU Time (hours)	120.98h	42.80h	41.52h
System Time (minutes)	70.02m	34.43m	29.04m

Input Data (L0):

Raw data from AIRS, AMSU-A1, AMSU-A2 instruments, and data about the spacecraft.

Output Data (L2):

Retrieved physical variables, e.g. temperature, humidity and ozone profiles, total precipitable water, cloud top height.





Hardware Performance Analysis

	Local Real Linux box (s4pt)	Nebula virtual Linux Box
Hardware	DELL PowerEdge 6800 with Dual-Core Xeon Processor 7100 series, 4 CPU	DELL PowerEdge c2100 with Quad-Core Xeon Processor 5500 series, 2 CPU
CPU (GHz)	8 cores * 3.16	16 cores * 2.8
Microarchitecture	65nm NetBurst	45nm Nehalem / 32nm Westmere

	NetBurst Microarchitecture	Nehalem/Westmere Microarchitecture
Cache L3	N/A	2 MB/core
FSB	Dual Independent 800MHz	QPI=6.4GT/s (Quick Path Interconnections)
Memory	DDR-2 400 ECC SDRAM (double channel)	DDR-3 (triple channel)

Netburst (65nm) --> **Core** (65nm) / **Penryn** (45nm) --> **Nehalem** (45nm)/**Westmere** (32nm)

Core = **2.5 x** NetBurst

Penryn = **1.8 x** Core

Nehalem/Westmere = **(1.2-2.0) x** Penryn.

➡ Nehalem/Westmere = **(5.4-9.0) x** NetBurst



Performance Analysis -1



Two-day AIRS L2 Processing at Nebula box and Local box

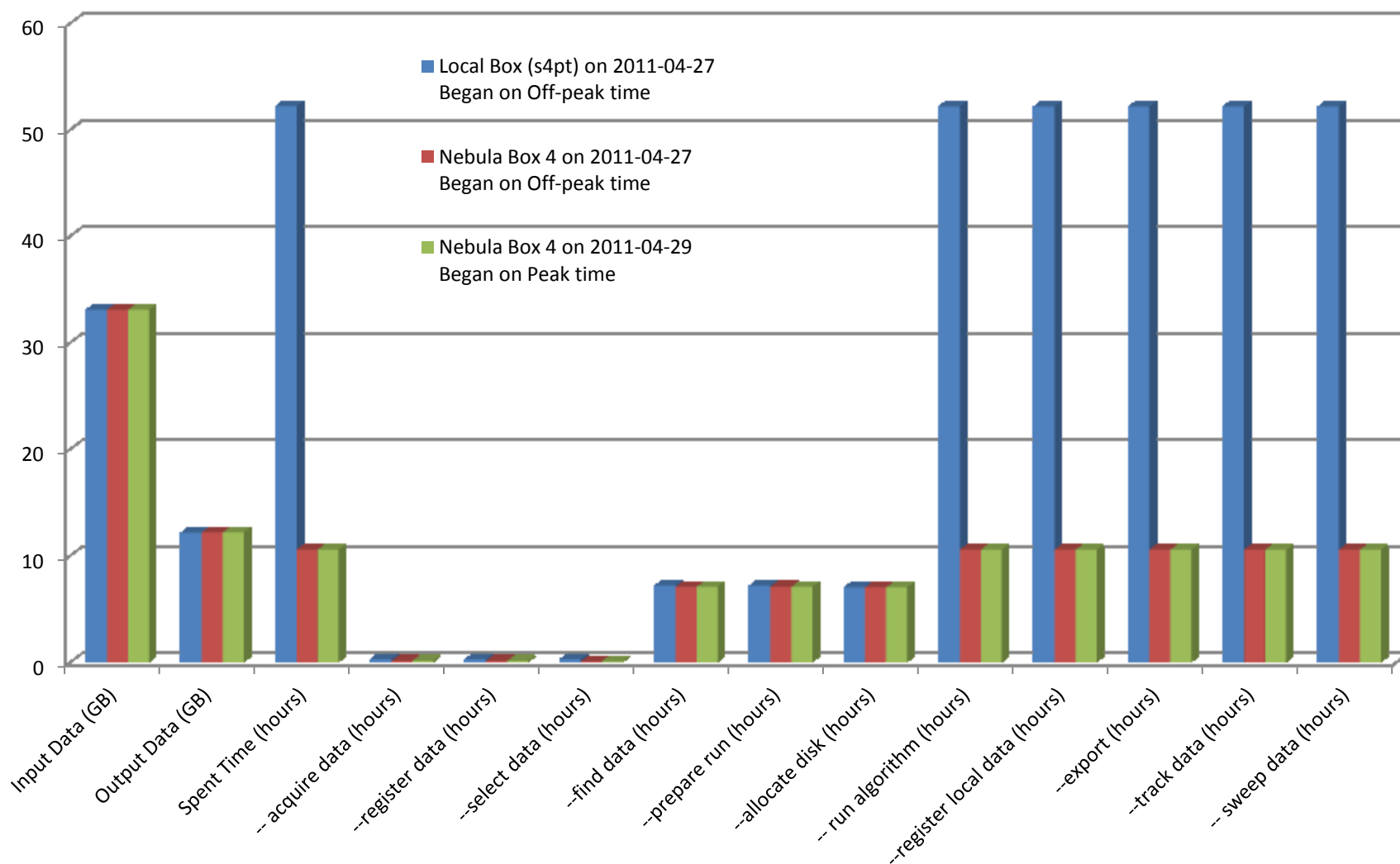
	Local Box On Off-peak time	Nebula Box 4 On Off-peak time	Nebula Box 4 On Peak time
Data Date	2010.123-124 (2010.05.02-03)	2010.123-124 (2010.05.02-03)	
Input Data	~33.1GB	~33.1GB	~33.1GB
Output Data	~12.16 GB	~12.2GB	12.2~ GB
Spent Time	52h 11m 58s	10h 34m 50s	10h 33m 43s
-- acquire data	17m 14s (21:36:33 - 21:53:47)	6m 33s (14:01:25 - 14:07:58)	6m 34s (07:58:32 - 08:05:06)
--register data	17m 21s (21:36:35 - 21:53:56)	6m 32s (14:01:28 - 14:08:00)	6m 43s (07:58:34 - 08:05:17)
--select data	16m 5s (21:36:36 - 21:52:41)	4m 26s (14:01:31 - 14:06:57)	5m 36s (07:58:37 - 08:04:13)
--find data	7h 10m 58s (21:36:37 - 4-28 04:47:35)	7h 5m 12s (14:01:34 - 21:06:46)	7h 4m 54s (07:58:40 - 15:03:34)
--prepare run	7h 10m 59s (21:36:38 - 04-28 04:47:37)	7h 5m 18s (14:01:37 - 21:06:55)	7h 4m 56s (07:58:43 - 15:03:39)
--allocate disk	7h 1m 39s (21:46:16 - 04-28 04:47:55)	7h 1m 50s (14:05:26- 21:07:16)	7h 1m 13s (08:02:32 - 15:03:45)
-- run algorithm	52h 11m 34s (04-27 21:36:40 - 04-30 01:48:14)	10h 34m 23s (14:01:40 - 04-28 00:36:03)	10h 34m 15s (07:58:46 - 18:32:01)
--register local data	52h 10m 40s (21:36:41 - 04-30 01:48:21)	10h 34m 27s (14:01:44 - 04-28 00:36:11)	10h 33m 16s (07:58:49 - 18:32:05)
--export	52h 10m 40s (21:36:42 - 04-30 01:48:22)	10h 34m 23s (14:01:46 - 04-28 00:36:09)	10h 33m 17s (07:58:52 - 18:32:09)
--track data	52h 10m 47s (21:36:44 - 04-30 01:48:31)	10h 34m 25s (14:01:50 - 04-28 00:36:15)	10h 33m 15s (07:58:55 - 18:32:10)
-- sweep data	52h 10m 40s (21:36:46 - 04-30 01:48:25)	10h 34m 1s (14:02:12 - 04-28 00:36:13)	10h 33m 10s (07:59:05 - 18:32:15)



Performance Analysis -2



Two-day AIRS L2 Processing at Nebula box and Local box



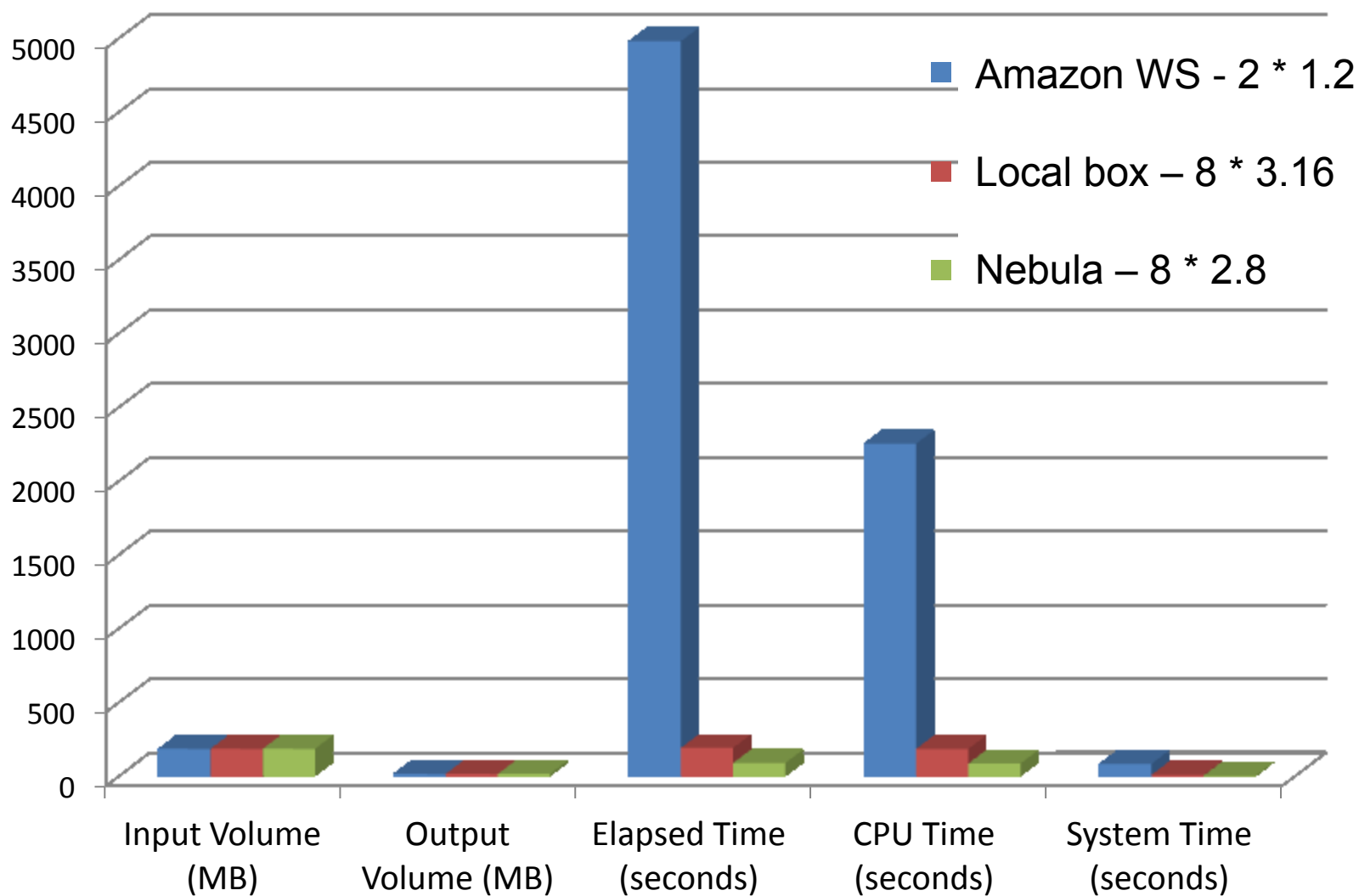


PREPQC 2011-111 (one-day)

One day (2011.111)	Amazon WS		Local Linux box	Nebula
Hardware Information	t1.micro : 613MB RAM Up to 2 * 1.2GHz Nehalem-based processor		16GB RAM, 4 * 3.16GHz NetBurst-based dual-core processor (8 cores)	8GB RAM, 2 * 2.8GHz Nehalem-based quad-core processor (8 cores)
Input Volume (MB)	184.67		184.67	184.67
Output Volume (MB)	14.89		14.95	15.01
Elapsed Time (seconds)	4980.25	5745.15	191.05	88.83
CPU Time (seconds)	2253.23	2395.79	184.21	85.51
System Time (seconds)	82.29	99.42	7.58	3.16



Performance comparison: AWS, Nebula, and local





Cost Estimation and Comparison -1



-- Nebula charge policies

❖ CPU charges

- \$0.12 per CPU-hr.
- \$0.48 per hour for an m1.large instance, which uses 4 CPUs.
- The charge applies whenever an instance is running, whether or not it is processing.
- Cloud applications should be designed to terminate non-processing instances wherever possible.

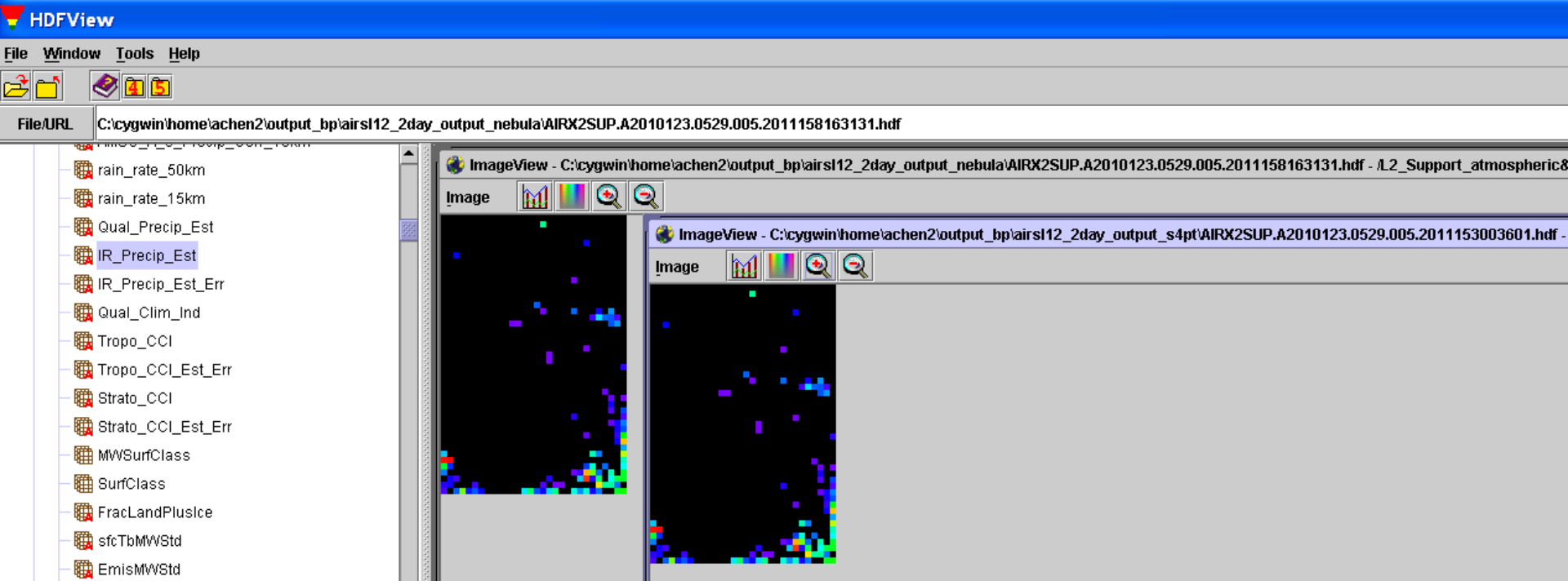
❖ Storage charges

- \$0.15 per GB-month apply to Volume storage and to Object Store storage.
- No charge for internal storage which comes within an instance (100GB) .
- Nebula does not charge for the storage used for your images themselves.
- Nebula does not charge for inputs and outputs, puts and gets, or network bandwidth usage.



Output Verification -1

AIRX2SUP (AIRS L2 Products)

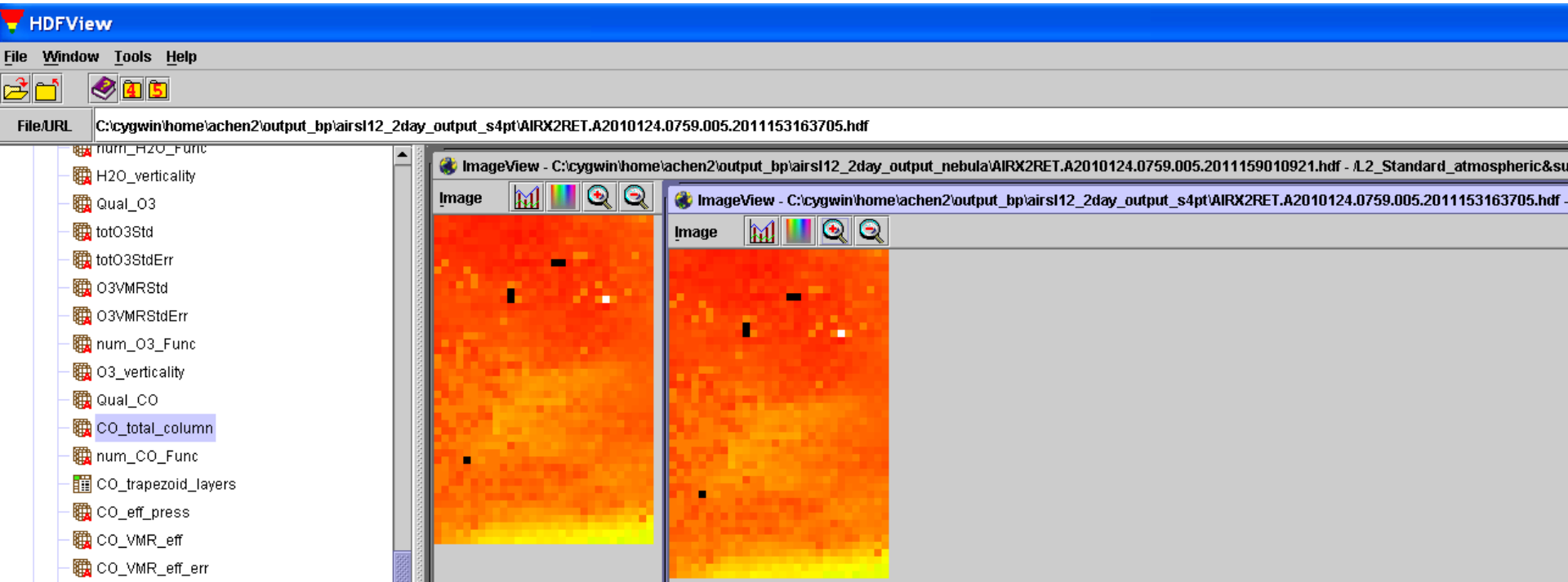




Output Verification -2



AIRX2RET (AIRS L2 Products)





Advantages -1



- ❖ User **Friendly** interface, access to and manage Nebula resources
 - **dashboard**: simple and convenient web interface
 - **Euca2ools**: fast and powerful command line tools
- ❖ **Better Performance**, compared with local box (details in appendix C)
- ❖ **Lower cost**, only pay for used time and resources (details in appendix C)
- ❖ **Scalability**, on-demand provisioning of resources in near real time and without users involvement for peak loads.
- ❖ **Cloning**, simple bundling process to save a modified/improved image. This is an excellent feature to **maintain**, **back up**, and **mirror** the systems; hence, increasing **reliability**.



Advantages -2



- ❖ VPN-based **high security** (1024 bit private and public key and X509 Cert.), **easy login** using private keys.
- ❖ **Knowledge base**:
 - Detailed how-to instructions for using Nebula via Dashboard and Euca2ools.
 - Fairly comprehensive FAQ, covering most common questions.
 - Helpful tutorial video for getting started.
- ❖ **Nebula Forum**, good venue for additional materials, user encountered bugs, solutions, and discussion.
- ❖ **Nebula team support**, responsive and eager to help; prompt response to general questions and resolving commonly encountered problems.



Challenges -1



❖ Stability

- **Instances** are not stable, operational access maybe lost and instances have to be rebooted. Before rebooting an instance, all attached volumes have to be detached.
- **Network** (FTP/wget) between Nebula and local machines is slow and not stable. Complications may arise from users attempting to ssh into Nebula instances during data transfers via FTP/wget (e.g. login failure, frequent FTP timeout, and throughput stalls).

❖ Under Developed

- Object Store not yet available
- Lack of tools for managing and monitoring running instances (e.g. Elastic Load Balancing, CloudWatch, Auto Scaling, etc.).

❖ Images, Volumes & Bundles

- Bare-bone images lacking trivial software packages (e.g. gcc, x11).
- When volume is attached, the specified location maybe not necessarily correspond to the entered location (e.g. /dev/vdh may end up as /dev/vdg).
- Any defects in the image you start with will be bundled up with your instance into your resulting image. (Defects in CentOS images result in bundling issues).



Challenges -2



❖ Gaps in Knowledge Base

- Lack of information on Nebula provided images.
- No troubleshooting tools.
- Details on **hardware** and **basic software** used by Nebula?

❖ Communication Concerns

- **Telecon:** Nebula used to have periodic telecon for end users to discuss problems needs, defects. These would be beneficial if they would return
- **Technical support:** Faster and more efficient technical support is needed
- **Forum:** Turn around for technical questions is long. Some posts are not responded to

❖ Size Limitation

- Instances: Maximum of 5 instances per project
- Volumes: 100GB volume storage per project (**exceptions can be requested directly*)
- CPUs: 16 cores.

❖ Commercial Software

- Uncertainty about 3rd-party commercial software installation on Nebula (e.g. licenses issues using instances with 3rd-party software in other projects, etc.).



Summary



- Three applications were successfully migrated to Nebula, including S4PM, AIRS L1/L2 algorithms, and Giovanni MAPSS.
- Nebula has some advantages compared with local machines (e.g. performance, cost, scalability, bundling, etc.)
- Nebula still faces some challenges (e.g. stability, object storage, networking, etc.).
- Migrating applications to Nebula is **feasible** but **time consuming**.
- Lessons learned from our Nebula experience will benefit future Cloud Computing efforts at GES DISC.



Team Members



Long Pham

Aijun Chen

Steven Kempler

Christopher Lynnes

Michael Theobald

Esfandiari Asghar

Jane Campino

Bruce Vollmer



Thank You for your attention !

Any Questions ?